Fairness-Aware Machine Learning for Social Bias Detection in Healthcare Research Datasets

Precious Kolawole PRECIOUSKOLAWOLE@CMAIL.CARLETON.CA
School of Computer Science, Carleton University, Canada

Abstract

This work presents an automated tool for detecting and measuring bias in healthcare datasets and predictive models. We evaluated fairness at both the data and algorithmic levels using metrics including Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), and Demographic Disparity. Using the SyntheticMass (healthcare expenses) and Brain Stroke healthcare datasets, we found that SyntheticMass showed substantial demographic imbalance (83.6% White patients) and age-based disparities (SPD: 0.82 for younger vs. elderly patients). While the Brain Stroke dataset exhibited more balanced demographics, we identified substantial disparities in stroke outcomes between age groups. Across both datasets, neural networks consistently outperformed traditional machine learning models on fairness metrics. In the Brain Stroke dataset, neural networks achieved both higher accuracy (94.8% vs. 91.8% for the best traditional model) and nearly perfect fairness scores (SPD: 0.000–0.0007; EOD: 0.000–0.0128). Additionally, we introduced a combined scoring metric that equally weights accuracy and fairness, providing researchers with a practical framework for model selection that prioritizes both dimensions. The interactive visualization dashboard makes fairness analysis accessible to medical researchers without specialized knowledge of fairness-aware machine learning.

Keywords: healthcare bias, fairness evaluation, machine learning, neural networks, statistical parity, equal opportunity, demographic disparity.

1. Introduction

Artificial intelligence (AI) is quickly becoming central in healthcare. Machine learning algorithms can help doctors diagnose diseases, suggest treatments, or predict which patients need urgent care. As exciting as this is, there is a hidden risk: If the AI is trained on biased or incomplete data, it can make unfair decisions that affect people's lives. There have already been real-world cases that show this danger. For example, a 2019 study found that an algorithm from a US hospital gave white patients better access to care than black patients (Obermeyer et al., 2019). In another case, an AI hiring tool at Amazon was biased against women (Dastin, 2018). There have even been AI systems that miss early signs of disease in seniors because too few elderly patients are included in their training data (Institute of Medicine Committee on the Future Health Care Workforce for Older Americans, 2008). Researchers creating new AI for healthcare might not realize when their datasets are missing important groups, or when their models treat some groups less fairly than others. By the time these biases appear in real-world decisions, it can be too late (or too complicated) to fix them. Our goal is to help researchers spot and measure these hidden biases early, before deployment.

In this work, we focus on two fundamental questions: (1) How can researchers identify and quantify data-level and algorithmic biases in healthcare datasets? (2) How do different machine learning approaches, especially neural networks versus traditional models, compare when it comes to balancing fairness and accuracy?

1.1. Related Prior Work

Many tools and studies have aimed to detect or highlight bias in machine learning, but few offer an all-in-one practical solution for healthcare researchers. For example, Intuit's `bias-detector` (Mishraky et al., 2022) uses metrics like Statistical Parity Difference and Equal Opportunity to detect gender and race bias in binary classifiers, but it relies on features (like names and zip codes) most relevant to US data, and users must customize it for broader applications. News Bias Detector (Raza et al., 2024) uses natural language models to spot political bias in text, but cannot handle structured healthcare datasets. Other tools, like AlgorithmAudit's Bias Detection (Algorithm Audit, 2023), use clustering to find underperforming user groups without needing protected attribute labels. However, they focus only on performance gaps, not on well-defined fairness metrics.

Within healthcare, most research has focused on identifying when bias has already occurred. For instance, Siddique et al. (Siddique et al., 2024) showed how clinical algorithms can accidentally deepen racial disparities, while Obermeyer et al. (Obermeyer et al., 2019) demonstrated how cost-based predictions led to underestimating the needs of black patients. While these studies are critical, they do not provide step-by-step tools for everyday researchers to check their own data and models for bias.

Our work was motivated by this gap: (1) We focus specifically on healthcare, (2) We integrate both data-level (who is in your dataset?) and algorithm-level (how fair are your predictions?) analyses in one easy-to-use tool, (3) We provide direct comparisons between traditional and neural network models on fairness as well as accuracy, and (4) We build an interactive tool that any researcher can use, even without expertise in fairness-aware machine learning.

1.2. Objectives

Our main objective is to create a practical and accessible tool that helps researchers find and reduce bias in healthcare data, from the very first stages of a project.

Specifically, our tool: (1) **Detects data-level bias** - like class imbalance and differences in positive or negative outcomes between demographic groups. (2) **Assesses algorithmic bias** - by measuring whether trained machine learning models (traditional or neural networks) have consistent performance and fairness across protected groups. (3) **Implements industry-standard fairness metrics**, including Statistical Parity Difference (IBM, 2023), Equal Opportunity Difference, Demographic Disparity (Amazon Web Services, 2023), and Average Odds Difference. (4) **Visualizes results in an interactive interface**, so users can instantly see and explore where bias may be lurking. By giving researchers the ability to investigate, understand, and resolve bias in their datasets and models, our tool supports more equitable AI in healthcare, helping to avoid harm before it happens.

1.3. Key Definitions and Concepts

We have defined the main concepts used throughout our paper. These terms often have precise mathematical meanings in fairness research:

- **Protected Attribute**: A feature or characteristic (such as race, gender, or age) that is legally or ethically important. Fairness analysis checks if outcomes for groups defined by protected attributes differ more than they should.
- **Reference Group**: The demographic group used as a baseline when calculating fairness metrics. This is typically the majority group in each protected attribute (e.g., the most common racial group). All fairness comparisons are made relative to this reference group.

- **Data-level Bias**: Imbalances or inequalities in the dataset itself, before you train any model. For example, if one group (like older adults) only makes up 5% of your data, or if the positive outcome rate is much higher for one gender than another.
- **Algorithmic Bias**: Unequal performance or predictions made by a trained model across groups. For example, if the model makes more mistakes for rural patients than for urban patients, even if both are present in the training data.
- **Statistical Parity Difference (SPD)**: The difference in the rate of favorable outcomes (such as being recommended for treatment) between two groups. Mathematically:

$$\text{SPD} = P(\text{favorable outcome}|\text{Group A}) - P(\text{favorable outcome}|\text{Group B})$$

- **Equal Opportunity Difference (EOD)**: The gap in true positive rates (i.e., sensitivity) between groups. This metric highlights whether one group is less likely to receive a correct positive prediction than another.

$$\text{EOD} = \text{TPR}_{\text{Group A}} - \text{TPR}_{\text{Group B}}$$

- **Average Odds Difference (AOD)**: The average of the difference in true positive rates and false positive rates between groups. AOD captures both the fairness in giving correct positive predictions and the mistakes where negatives are falsely predicted as positive.

$$\text{AOD} = 0.5 \times [(\text{FPR}_{\text{Group A}} - \text{FPR}_{\text{Group B}}) + (\text{TPR}_{\text{Group A}} - \text{TPR}_{\text{Group B}})]$$

- **Demographic Disparity (DD)**: The difference between a group's share of negative (or positive) outcomes and their share of the overall population. If a group makes up 10% of the data but receives 30% of negative outcomes, there is demographic disparity.

$$\text{DD} = P(\text{Group}|\text{Unfavorable/favorable}) - P(\text{Group}|\text{Population})$$

- **Class Imbalance**: When one group or outcome category is much more common than others.

2. Methods

Our Social Bias Detection tool systematically evaluates biases in healthcare datasets through automated data-level and algorithmic-level analyses. We designed the tool using a Streamlit-powered user interface where users upload datasets in CSV format. All subsequent pre-processing, analysis, and visualization steps are automated.

2.1. Datasets

We conducted experiments on two synthetic but realistic healthcare datasets to demonstrate generalizability and robustness.

SyntheticMass Dataset: This open-source collection of simulated patient records (Walonoski et al., 2018), based on Massachusetts demographics, includes the `patients.csv` file with 12,353 records containing demographic features (race, ethnicity, gender, birthdate) and healthcare-related metrics. We selected `HEALTHCARE EXPENSES` as the target variable for fairness analysis.

Brain Stroke Dataset: Containing 4,981 patient records (Pathan et al., 2020), this dataset includes clinical and demographic information such as gender, age, hypertension, heart disease status, marital status, work type, residence type, glucose levels, BMI, and smoking status. The target variable is binary stroke occurrence (1 = stroke, 0 = no stroke).

2.2. Preprocessing Pipeline

Both datasets underwent consistent preprocessing steps:

1. **Missing Value Handling:** Missing data were imputed using mode for categorical variables and median for numerical variables. The Brain Stroke Dataset contained no missing values.

2. **Feature Engineering:** Age features were identified and processed: in the Brain Stroke dataset, age is numerical, while in SyntheticMass, age was derived from birthdate. We applied binning (Google Developers, 2023) to create age groups (0–18, 19–35, 36–50, 51–65, and 65+).

3. **Target Variable Preparation:** For continuous variables like healthcare expenses in SyntheticMass, binary targets were created by comparing values to the median; values above the median were classified as unfavorable outcomes (high expenses). In the Brain Stroke dataset, the `stroke` attribute is already binary (1 for stroke, 0 for no stroke); for fairness calculations, we ensured consistent encoding so that 1 always indicates the favorable outcome.

4. **Categorical Encoding:** LabelEncoder from scikit-learn was used to convert categorical protected attributes to numerical format for all models, maintaining reference mappings for interpretation.

2.3. Data-Level Bias Analysis

Data-level bias analysis examines the dataset before any model training. We implemented three complementary approaches:

Class Imbalance Assessment: We calculated the percentage representation of each group within protected attributes (race, gender, age group, etc.), flagging groups with less than 10% of the dataset as underrepresented (Bellamy et al., 2018).

Statistical Parity Difference (SPD): Using the formal definition from Section 1.3, we calculated SPD for each protected attribute by comparing the favorable outcome rate in the reference group (typically the most prevalent group) to those in other groups. Disparities were categorized as Minimal ($|\text{SPD}| < 0.05$), Small ($0.05 \leq |\text{SPD}| < 0.10$), or Substantial ($|\text{SPD}| \geq 0.10$) (Bellamy et al., 2018).

Demographic Disparity (DD): Following Section 1.3, we computed DD for each group as the difference between its proportion among unfavorable (or favorable) outcomes and its proportion in the full dataset, again classifying results by the same thresholds as for SPD.

2.4. Algorithm-Level Bias Analysis

We analyzed model-level fairness by splitting each dataset into training (70%) and test (30%) sets, with a fixed random seed of 42 to ensure reproducibility. Protected attributes were retained in the feature set to evaluate their possible influence on prediction.

Traditional Machine Learning Models: Using LazyPredict, we trained and evaluated over 20 classification algorithms, selecting the top five based on accuracy, ROC AUC, and F1 scores.

Neural Network Architectures: Three architectures of varying complexity were implemented: (1) Simple Network - a single hidden layer with 16 neurons, ReLU activation; (2) Deep Network - four hidden layers (64, 32, 16, and 8 neurons), ReLU activation, dropout rate 0.3; (3) Residual Network - two residual blocks with skip connections, each containing two dense layers (32 neurons) and dropout rate 0.2. All neural networks used binary cross-entropy loss, Adam optimizer (learning rate: 0.001), early stopping with patience of 5 epochs, batch size 32, and up to 10 epochs of training.

For each trained model, we calculated the percentage of correct predictions on the test set (Accuracy), the difference in favorable rates across protected groups (SPD), the difference in true positive rates across protected groups (EOD) and the average difference in false positive and true positive rates across protected groups (AOD). For algorithmic fairness, we computed these metrics using model predictions on the test set rather than the actual labels used in data-level analysis. For example, the algorithmic SPD is calculated as:

$$\text{SPD}_{\text{alg}} = P(\hat{Y} = 1|\text{Group A}) - P(\hat{Y} = 1|\text{Group B})$$

where $\hat{Y}$ denotes the model's predictions.

2.5. Fairness Scoring System

To compare models on both accuracy and fairness, we developed a combined scoring system:

1. Accuracy was normalized by dividing by the maximum accuracy among all models.

$$\text{Accuracy}_{\text{normalized}} = \frac{\text{Accuracy}}{\text{Accuracy}_{\text{max}}}$$

2. A fairness score was computed by normalizing and averaging SPD and EOD:

$$\text{Fairness score} = 1 - \frac{|\text{SPD}|}{|\text{SPD}_{\text{max}}|} - \frac{|\text{EOD}|}{|\text{EOD}_{\text{max}}|}$$

   where $|\text{SPD}_{\text{max}}|$ and $|\text{EOD}_{\text{max}}|$ are the maximum absolute SPD and EOD values observed across all models.

3. The combined score was computed as (a higher combined score indicates better trade-off between predictive performance and fairness):

$$\text{Combined score} = 0.5 \times \text{Accuracy} + 0.5 \times \text{Fairness score}$$

2.6. Validation and Reliability

To ensure analytic robustness and reproducibility, we cross-validated our fairness metric implementations against IBM's AI Fairness 360 toolkit (Bellamy et al., 2018), using both `BinaryLabelDatasetMetric` and `ClassificationMetric`. For statistical reliability and unbiased model comparisons, all machine learning models, including neural networks, were freshly initialized for each run, with no parameter sharing across experiments.

2.7. User Interface and Accessibility

We designed our tool to be accessible to researchers without specialized expertise in fairness-aware machine learning (see Figures 1 and 2). The complete implementation, including source code and a demonstration video, is available in our research repository: https://github.com/precillieo/social-bias-detection-tool.

3. Results

Below are our findings from the SyntheticMass and Brain Stroke Healthcare datasets. For each, we report data-level bias metrics followed by algorithmic fairness results.

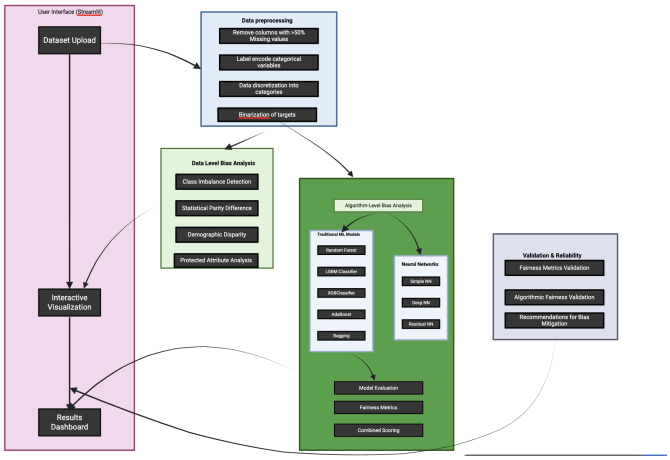

Figure 1: Architectural overview of the Social Bias Detection framework.

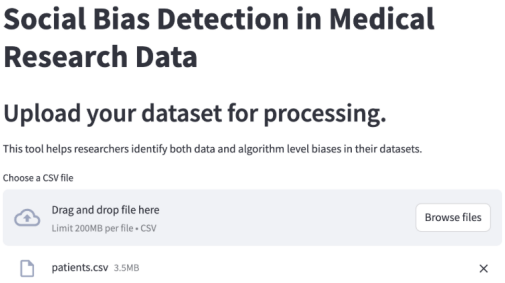

Figure 2: Interactive user interface of the Social Bias Detection tool.

3.1. Data-Level Bias Analysis

3.1.1. Class Imbalance

Table 1 presents the demographic distributions across both datasets. The SyntheticMass dataset exhibited substantial racial imbalance (83.6% White), while the Brain Stroke dataset showed moderate gender imbalance (58.4% Female) and marital status imbalance (65.9% Married). Both datasets showed contrasting age distributions: SyntheticMass skewed toward older adults (33.1% in 65+ category), while the Brain Stroke dataset had more balanced age representation (17.7%–22.8% across categories).

3.1.2. Statistical Parity Difference (SPD)

For each protected attribute, the SPD results are summarized side-by-side in Table 2. SPD is reported for every group (relative to the reference group listed). In the SyntheticMass dataset, substantial age-based disparities were observed, with much higher SPD values for the younger groups (0–18 and 19–35, both 0.8168) relative to the 65+ reference group. In contrast, most SPD values for ethnicity and gender were minimal or small. In the Brain Stroke dataset, SPD values were generally close to zero (minimal), with the exception of the 65+ age group (SPD: 0.1009, classified as substantial) compared to the 51–65 reference group. No other group exhibited substantial SPD.

3.1.3. Demographic Disparity (DD)

Table 3 compares Demographic Disparity metrics across both datasets. In the SyntheticMass dataset, DD values revealed major age-based skew: the 65+ group was overrepresented among unfavorable outcomes (DD: 0.2099), while the 19–35 and 0–18 groups were underrepresented (DD: -0.1915 and -0.1184, respectively). By comparison, no group in the Brain Stroke dataset exceeded the minimal threshold for demographically based disparity, with all DD values remaining close to zero across age, work type, and other attributes.

3.2. Algorithm-Level Fairness Analysis

We evaluated multiple model architectures for predicting high healthcare expenses in the SyntheticMass dataset. Table 4 reports accuracy and fairness metrics for each model. Traditional

Table 1: Class Imbalance Analysis: Demographic Distributions in Both Datasets

SyntheticMass Dataset			**Brain Stroke Dataset**		
Demographic	**Count**	**Percentage**	**Demographic**	**Count**	**Percentage**
	Race			*Gender*	
White	10,328	83.6%	Female	2,907	58.4%
Black	1,100	8.9%	Male	2,074	41.6%
Asian	842	6.8%			
Native	73	0.6%		*Marital Status*	
Other	9	0.1%	Yes	3,280	65.9%
			No	1,701	34.1%
	Ethnicity			*Work Type*	
Non-Hispanic	11,036	89.3%	Private	2,860	57.4%
Hispanic	1,316	10.7%	Self-employed	804	16.1%
			Children	673	13.5%
	Gender		Govt_job	644	12.9%
Female	6,253	50.6%			
Male	6,099	49.4%		*Residence Type*	
			Urban	2,532	50.8%
	Age Group		Rural	2,449	49.2%
65+	4,093	33.1%			
51-65	2,434	19.7%		*Age Group*	
19-35	2,366	19.2%	51-65	1,134	22.8%
36-50	1,997	16.2%	36-50	1,047	21.0%
0-18	1,462	11.8%	19-35	961	19.3%
			65+	959	19.3%
			0-18	880	17.7%

Table 2: Statistical Parity Difference (SPD) values by protected attribute and group for SyntheticMass (left) and Brain Stroke (right).

SyntheticMass Dataset

Attribute	**Group**	**Reference**	**SPD**
Race	Asian	White	0.0529
	Black	White	-0.0209
	Native	White	-0.0050
	Other	White	0.1685
Ethnicity	Nonhispanic	Hispanic	0.0272
Gender	Male	Female	-0.0539
Age Group	0–18	65+	0.8168
	19–35	65+	0.8168
	36–50	65+	0.3506
	51–65	65+	0.0353

Brain Stroke Dataset

Attribute	**Group**	**Reference**	**SPD**
Gender	Male	Female	0.0039
Ever Married	No	Yes	-0.0497
Work Type	Self-empl.	Private	0.0291
	Govt_job	Private	-0.0005
	Children	Private	-0.0488
Residence	Rural	Urban	-0.0072
Age Group	65+	51–65	0.1009
	36–50	51–65	-0.0436
	19–35	51–65	-0.0607
	0–18	51–65	-0.0595

models achieved higher accuracies (up to 0.899), but showed poorer fairness across both SPD (0.4080–0.4455) and EOD (0.8050–0.8496). Neural network models yielded slightly lower accuracy (0.8430–0.8492), but achieved lower SPD (0.3775–0.4277) and EOD (0.7300–0.7763) values, result-

Table 3: Demographic Disparity Comparison Between Datasets

SyntheticMass Dataset			Brain Stroke Dataset		
Group	**DD**	**Assessment**	**Group**	**DD**	**Assessment**
Age Groups			*Age Groups*		
65+	0.2099	Substantial	0-18	0.0088	Minimal
51-65	0.1109	Substantial	19-35	0.0099	Minimal
36-50	-0.0109	Minimal	36-50	0.0070	Minimal
19-35	-0.1915	Substantial	51-65	-0.0029	Minimal
0-18	-0.1184	Substantial	65+	-0.0229	Minimal
Race			*Work Type*		
White	0.0031	Minimal	Private	-0.0012	Minimal
Black	0.0040	Minimal	Self-employed	-0.0053	Minimal
Asian	-0.0070	Minimal	Children	0.0067	Minimal
Native	0.0001	Minimal	Govt_job	-0.0002	Minimal
Other	-0.0002	Minimal			

ing in higher combined scores. The ResidualNN model provided the best trade-off, achieving a combined score of 0.5448.

Table 4: Algorithmic fairness evaluation for SyntheticMass dataset

Model	**Category**	**Accuracy**	**SPD**	**EOD**	**Combined Score**
ResidualNN	Neural Network	0.8478	0.3775	0.7300	**0.5448**
DeepNN	Neural Network	0.8492	0.3902	0.7444	0.5342
AdaBoostClassifier	Traditional	0.8907	0.4080	0.8050	0.5295
RandomForestClassifier	Traditional	0.8991	0.4347	0.8408	0.5087
BaggingClassifier	Traditional	0.8886	0.4355	0.8309	0.5053
LGBMClassifier	Traditional	0.8988	0.4425	0.8485	0.5018
SimpleNN	Neural Network	0.8430	0.4277	0.7763	0.5003
XGBClassifier	Traditional	0.8969	0.4455	0.8496	0.4988

For the Brain Stroke dataset, we observed different performance characteristics, as summarized in Table 5. Neural network models achieved both higher accuracy (0.947–0.948) and near-perfect fairness metrics (SPD: 0.000–0.0007, EOD: 0.000–0.0128), resulting in the highest combined score (up to 0.9996). In contrast, traditional models had either lower accuracy (down to 0.708) and/or higher group disparities (SPD: 0.003–0.328, EOD: 0.000–0.846), leading to substantially lower combined scores.

Table 5: Algorithmic fairness evaluation for Brain Stroke dataset

Model	**Category**	**Accuracy**	**SPD**	**EOD**	**Combined Score**
DeepNN	Neural Network	0.9478	0.0000	0.0000	**0.9996**
ResidualNN	Neural Network	0.9478	0.0000	0.0000	0.9996
SimpleNN	Neural Network	0.9485	0.0007	0.0128	0.9957
Perceptron	Traditional	0.9445	0.0033	0.0000	0.9953
BernoulliNB	Traditional	0.9177	0.0595	0.2821	0.8550
QuadraticDA	Traditional	0.8749	0.1278	0.5256	0.7085
GaussianNB	Traditional	0.8562	0.1572	0.6282	0.6458
NearestCentroid	Traditional	0.7084	0.3278	0.8462	0.3734

4. Discussion

Our fairness detection tool provided actionable insights into bias within healthcare datasets and predictive models. These results help answer a fundamental question: *How fair is my data and my model, and how can I know?* We recommend using the following guideline: Statistical Parity Differences (SPD) above 0.10 (Bellamy et al., 2018) typically indicate substantial bias that may require further attention. A dataset or model can be considered "fair" if all group-wise metrics are consistently minimal or small, as indicated in our tables and visual outputs. Importantly, not all disparities signal unfairness; some reflect real clinical patterns. For example, in SyntheticMass, high SPD values for younger age groups likely reflect true lower healthcare expenses for youth compared to older adults. Domain knowledge remains essential for interpreting which disparities are concerning. In our algorithmic analysis, neural networks consistently offered a better balance of fairness and accuracy than traditional models, particularly in the Brain Stroke dataset, where neural networks achieved SPD and EOD near zero. We attribute this to the ability of neural networks to model complex, group-dependent patterns, even with some imbalance. However, in SyntheticMass, where class imbalance is severe (e.g., 84% White patients), the fairness difference between neural and traditional models diminished. Thus, neural networks might improve equity, but only when the training data have adequate group representation.

As future work, our combined accuracy-fairness scoring could be improved with adaptive weighting tuned to clinical priorities. Adding explainable AI methods would also help users understand not just if bias exists, but why. All fairness and accuracy metrics in this study were fully validated and reflect corrected computations.

5. Conclusions

We developed and validated a comprehensive tool for detecting bias in healthcare datasets at both the data and algorithmic levels. Our methods assessed class imbalance, statistical parity difference, and demographic disparity, uncovering racial and age-based imbalances, especially in the SyntheticMass dataset. Across models, neural networks, particularly ResidualNN, struck the best balance between accuracy and fairness, outperforming traditional models on key fairness metrics. This highlights the importance of considering equity, not just predictive accuracy, when selecting models for healthcare AI. The tool's interactive dashboard makes fairness analysis accessible to all researchers, helping them identify and address potential biases with minimal technical overhead. Although more work is needed to extend validation and add interpretability features, our results show that this framework is a practical step toward more equitable, and trustworthy healthcare AI.

6. Acknowledgements

The author thanks Professor Matthew S. Holden for mentorship, insights, and valuable feedback on this work. We also thank Ilerioluwakiiye Abolade for insightful discussions and contributions on this work.

References

Algorithm Audit. Bias detection tool, 2023. URL https://algorithmaudit.eu/technical-tools/bdt/.

Amazon Web Services. Conditional demographic disparity - sagemaker developer guide, 2023. URL https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-data-bias-metric-cddl.html.

R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. Natesan Ramamurthy, J. T. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*, 2018. URL https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG. Accessed: 2024-07-10.

Google Developers. Binning - machine learning crash course, 2023. URL https://developers.google.com/machine-learning/crash-course/numerical-data/binning.

IBM. Statistical parity difference – ibm cloud pak for data documentation, 2023. URL https://dataplatform.cloud.ibm.com/docs/content/wsj/model/wos-fairness-stat-parity-difference.html.

Institute of Medicine Committee on the Future Health Care Workforce for Older Americans. *Retooling for an aging America: Building the health care workforce*. National Academies Press, 2008. URL https://www.ncbi.nlm.nih.gov/books/NBK215400/.

E. Mishraky, A. Ben Arie, Y. Horesh, and S. Meir Lador. Bias detection by using name disparity tables across protected groups. *Journal of Responsible Technology*, 9:100020, 2022. doi: 10.1016/j.jrt.2021.100020.

Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. doi: 10.1126/science.aax2342.

Muhammad Salman Pathan, Zhang Jianbiao, Deepu John, Avishek Nag, and Soumyabrata Dev. Identifying stroke indicators using rough sets. *IEEE Access*, 8:210318–210327, 2020. doi: 10.1109/ACCESS.2020.3039439.

S. Raza, M. Rahman, and M. R. Zhang. Beads: Bias evaluation across domains. *arXiv preprint arXiv:2406.04220*, 2024.

S. M. Siddique, K. Tipton, B. Leas, S. R. Greysen, N. K. Mull, M. Lane-Fall, C. A. Umscheid, and A. S. Navathe. The impact of health care algorithms on racial and ethnic disparities: A systematic review. *Annals of Internal Medicine*, 177(4):484–496, 2024. doi: 10.7326/M23-2960.

J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, and S. McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238, 2018. doi: 10.1093/jamia/ocx079.